Interspeech 2022
18-22 September 2022, Incheon, Korea

CAUSE: Crossmodal Action Unit Sequence Estimation from Speech

Hirokazu Kameoka, Takuhiro Kaneko, Shogo Seki, and Kou Tanaka

NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation
{hirokazu.kameoka.uh,takuhiro.kaneko.tb,shogo.seki.va,kou.tanaka.ef}@hco.ntt.co.jp

Abstract

This paper proposes a task and method for estimating a sequence of facial action units (AUs) solely from speech. AUs were introduced in the facial action coding system to objectively describe facial muscle activations. Our motivation is that AUs can be useful continuous quantities for representing speaker's subtle emotional states, attitudes, and moods in a variety of applications such as expressive speech synthesis and emotional voice conversion. We hypothesize that the information about the speaker's facial muscle movements is expressed in the generated speech and can somehow be predicted from speech alone. To verify this, we devise a neural network model that predicts an AU sequence from the mel-spectrogram of input speech and train it using a large-scale audio-visual dataset consisting of many speaking face-tracks. We call our method and model "crossmodal AU sequence estimation/estimator (CAUSE)". We implemented several of the most basic architectures for CAUSE, and quantitatively confirmed that the fully convolutional architecture performed best. Furthermore, by combining CAUSE with an AU-conditioned image-to-image translation method, we implemented a system that animates a given still face image from speech. Using this system, we confirmed the potential usefulness of AUs as a representation of non-linguistic features via subjective evaluations.
Index Terms: Non-linguistic features, action units (AUs), crossmodal transfer, facial animation synthesis

1. Introduction

Being able to quantify non-linguistic information in speech, such as expressions and moods can be useful for a variety of applications. These applications include dialogue systems, emotional voice conversion, expressive speech synthesis, and talking head generation.

One example of non-linguistic feature extraction from speech is the extraction of parameters related to the movement of the thyroid cartilage based on a physiologically grounded model for describing voice fundamental frequency contours (called the Fujisaki model) [1]. The Fujisaki model is particularly appealing in that its parameters characterize the intonation of speech in a continuous and physiologically meaningful way, and once these parameters can be extracted, it allows natural and detailed control of the intonation by manipulating them. However, the downside is these features are so low-order that it is not easy to interpret how they are related to higher-order information such as expressions and moods. Therefore, there are still many issues that need to be resolved before they can be used as an effective intermediate representation in the applications described above.

Another example is the extraction of the emotional content of speech. Despite recent advances in the field of speech emotion recognition, obtaining accurate and reliable techniques remains a formidable challenge, hampered by several obstacles. One such obstacle is the limited number of reliable datasets suitable for real-world applications. While there are several publicly available emotion-labeled datasets consisting of acted speech (which is rarely found in real-world conversation), there are very few when it comes to those consisting of naturally spoken or spontaneous speech. Another obstacle is the ambiguity of emotional labels. Since emotions are subjective, even if a labeled dataset could be created, it would suffer from low human annotator agreement. To overcome these obstacles, an approach called crossmodal transfer has recently been proposed [2]. The idea is to first train a facial expression recognizer using labeled facial images, and then train a speech emotion recognizer using the audio part of unlabeled videos of human speech in the wild so that its predictions match those of the facial expression recognizer applied to the face-tracks of the corresponding segments. The hypothesis behind this idea is that the emotional content of speech is correlated with the facial expression of the speaker, and that facial expression recognition is a relatively easier task to solve. Nevertheless, in pretraining the facial expression recognizer, the aforementioned problem of label ambiguity remains an issue. To address this, the authors propose to use an emotion-labeled dataset of still facial images called FERPlus [3]. Unlike other existing emotion-labeled datasets, FERPlus was created by incorporating a measure of uncertainty into the labeling scheme by having ten human annotators label each image. As a result, the trained speech emotion recognizer will be able to predict a sequence of the distributions over eight emotional states (neutral, happiness, surprise, sadness, anger, disgust, fear, and contempt) from speech. While this approach is indeed powerful for a speech emotion recognition task, emotional states defined in a categorical space would be too abstract to accurately represent the fine and subtle expressions and moods of speech.

Given the aforementioned applications in mind, the features we wish to extract from speech are quantities that are easy to interpret, like emotional states, but also capture fine and subtle features that are linked to the physiological or anatomical mechanisms behind speech, like the Fujisaki model parameters. As quantities that satisfy these requirements, in this paper, we focus on the facial action units (AUs) defined in the facial action coding system (FACS) [4]. AUs are facial muscular activity units that are related to the contraction or relaxation of specific facial muscles, and they can describe nearly any anatomically possible facial expression. In this paper, inspired by the idea of the abovementioned crossmodal transfer, we propose a task and a method for estimating a sequence of the AU intensities solely from speech. Since it is not clear how much AU-related information is actually contained in speech, an interesting question is whether it is really possible to estimate the AU intensities from speech alone, and if so, how precise the estimation can be. One of our goals is to gain some insight

10.21437/Interspeech.2022-11232

into this non-trivial question. We collectively refer to our task, method, and model as the "crossmodal action unit sequence estimation/estimator (CAUSE)". In addition, by combining the proposed CAUSE with "GANimation" [5], an AU-conditioned image-to-image translation method based on generative adversarial networks (GANs) [6, 7], we implement a system that can animate a given still face image from speech. This system allows us to visually check the subjective validity of the AUs extracted using CAUSE.

2. Related Work

In this paper, we choose to implement a system that generates facial animations from speech mainly for the purpose of subjective evaluation of CAUSE. While many attempts have already been made to implement speech-to-face systems, owing to space limitations, we mention a few that are relevant below.

Examples from the most recent studies include talking-head generation [8, 9], lip-syncing [10], and face image reconstruction from speech (Speech2Face) [11]. The first two focus mainly on predicting mouth and head movements and gestures from speech, while the last one focuses on predicting the appearance of the person speaking, such as the person's age, gender and ethnicity. To the best of our knowledge, no attempt has yet been made to utilize AUs for these tasks, and we believe that incorporating CAUSE would further augment these systems.

3. Method

3.1. AU Detection

AUs were introduced in the FACS to objectively describe facial muscle activations. While there are few freely available tools for AU detection, a toolkit called OpenFace[1] [12] provides a reliable function for detecting the presence and intensity of 18 AUs, either from images or from videos [13]. In our work, we used OpenFace to extract the intensities of 17 AUs[2] from an image or each frame of a video, excluding the AU for which only presence prediction was given. We used the pretrained models for all the functions made available by the developers of OpenFace.

3.2. Datasets

For the current task, we used the VoxCeleb2 dataset [14] for training and testing CAUSE, and the CelebA dataset [15] for training GANimation and testing facial animation synthesis using CAUSE and GANimation. The VoxCeleb2 dataset contains over one million utterances from 6,112 celebrities, extracted from videos uploaded to YouTube. The CelebA dataset consists of more than 200,000 celebrity images.

3.3. CAUSE

Our CAUSE is inspired by the idea of the crossmodal transfer method [2]. For each video excerpted from the VoxCeleb2 dataset, we first compute the mel-spectrogram from the audio part with a hop size matched to the video frame rate (25fps). By doing so, we can obtain a pair of face-track sequences and mel-spectrograms with each frame synchronized. We then perform AU detection using OpenFace on the face-track sequence to extract a sequence of AU intensities. With the above preparation, our goal is to train a CAUSE network f_θ that takes an 80-channel mel-spectrogram $\mathbf{X} \in \mathcal{R}^{80\times N}$ of length N as input and produces a sequence of the predicted AU intensities of the same length $\hat{\mathbf{Y}} \in \mathbb{R}^{17\times N}$

$$\hat{\mathbf{Y}} = f_\theta(\mathbf{X}). \tag{1}$$

By using $\mathbf{Y} \in \mathbb{R}^{17\times N}$ to denote the AU intensity sequence predicted using OpenFace from the corresponding face-track sequence, the training loss is defined by

$$L(\theta) = \mathbb{E}_{(\mathbf{X},\mathbf{Y})}[\|\mathbf{Y} - f_\theta(\mathbf{X})\|_1], \tag{2}$$

where $\mathbb{E}_{(\mathbf{X},\mathbf{Y})}[\cdot]$ means the sample mean over all training examples, and $\|\cdot\|_1$ denotes the mean of the absolute values of all the matrix elements. An illustration of CAUSE network training is shown in Fig. 1.

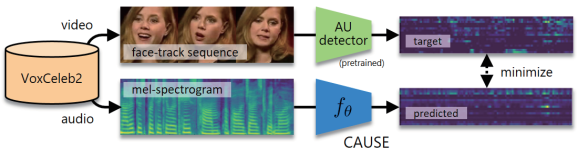

Figure 1: *Illustration of CAUSE network training*

The architecture of f_θ needs to be designed according to the range of the mel-spectrogram on which the AU intensity at each frame can depend. Since this is not evident, we tested the following four types of architectures for comparison: a frame-independent fully connected network (multilayer perceptron; MLP), a recurrent network (RNN), a regular convolutional network (CNN), and a dilated convolutional network (DCNN). If the AU intensity at each frame depends only on the mel-spectrum at the corresponding frame, then the MLP architecture should perform on par with the others. If not, a comparison of the remaining three should give us some indication of the range of the mel-spectrogram to focus on as a clue for inference. These architectures are devised as follows.

MLP: In the MLP architecture, the network consists of ten fully-connected linear layers with 128, 128, 64, 64, 32, 32, 16, 16, 8, and 17 output channels, respectively, each followed by a dropout layer with a dropout ratio of 0.1 and a leaky rectified linear unit (LReLU) function with a negative slope of 0.1 except for the first and last layers. Weight normalization is applied to all the weights. This architecture is designed to process the mel-spectrum of each frame independently.

RNN: In the RNN architecture, the network consists of a fully-connected linear layer with 128 output channels, a four-layer bidirectional long-short term memory (BiLSTM) network with 64 hidden units, and a fully-connected linear layer with 17 output channels. Weight normalization is applied to the first and last linear layers.

CNN: In the CNN architecture, the network consists of a fully-connected linear layer with 128 output channels, eight 1D convolution layers with a kernel size of 3 and with 128, 64, 64, 32, 32, 16, 16, and 8 output channels, respectively, each followed by a dropout layer with a dropout ratio of 0.1 and a gated linear unit (GLU) function, and a fully connected linear layer with

[1] https://github.com/TadasBaltrusaitis/OpenFace

[2] AU1: inner brow raiser. AU2: outer brow raiser. AU4: brow lowerer. AU5: upper lid raiser. AU6: cheek raiser. AU7: lid tightener. AU9: noise wrinkler. AU10: upper lip raiser. AU12: lip corner puller. AU14: dimpler. AU15: lip corner depressor. AU17: chin raiser. AU20: lip streched. AU23: lip tightener. AU25: lips part. AU26: jaw drop. AU45: blink.

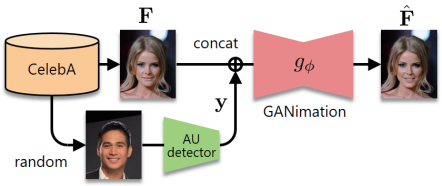

Figure 2: *Illustration of GANimation training*

17 output channels. Weight normalization is applied to all the learnable weights.

DCNN: The DCNN architecture is the same as the CNN architecture except that the dilation factors of the convolution layers are set to 1, 3, 9, 27, 1, 3, 9, and 27, respectively. With these settings, the receptive field of DCNN becomes much larger than that of CNN even with the same number of parameters. Relatively speaking, the CNN architecture is designed to look at a local region in detail, while the DCNN architecture is designed to look broadly at a large region.

3.4. GANimation

Once the AU intensity sequence has been extracted from speech, we use it to animate a given still facial image to "visualize" the expression and mood of speech. To this end, we choose to use GANimation [5], an image-to-image translation method that uses AUs as conditioning variables.

Let $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ be an input face image, where H and W are the vertical and horizontal sizes, respectively, and C is the channel number ($C = 3$ for RGB images). Let $\mathbf{y} \in \mathbb{R}^{17}$ be an AU intensity vector corresponding to a target expression. The generator g_ϕ is a neural network that takes as inputs $\mathbf{F}$ and $\mathbf{y}$ and ouputs a converted version of $\mathbf{F}$:

$$\hat{\mathbf{F}} = g_\phi(\mathbf{F}, \mathbf{y}), \tag{3}$$

with some architecture parameterized by ϕ. Below, we describe the architectural design of g_ϕ and the losses used for its training, along with the ideas behind them.

Architecture: g_ϕ is configured so that the attention mask $\mathbf{A} \in (0, 1)^{1\times H\times W}$ and color mask $\mathbf{C} \in \mathbb{R}^{C\times H\times W}$ are first generated as internal representations, and then $\hat{\mathbf{F}}$ is given as

$$\hat{\mathbf{F}} = (\mathbf{1} - \mathbf{A}) \odot \mathbf{C} + \mathbf{A} \odot \mathbf{F} \tag{4}$$

using $\mathbf{A}$, $\mathbf{C}$, and $\mathbf{F}$, where $\mathbf{1}$ represents an array with all 1 elements, and $\odot$ denotes broadcasting followed by elementwise multiplication. The coordinates where the attention mask takes a zero value indicate the area of the input image to be converted. The color mask is a quantity that corresponds to the difference between the converted and input images.

Image Adversarial Loss: In training, one of the goals is to make the generated image $\hat{\mathbf{F}}$ look as realistic as possible. Based on the Wasserstein GAN formulation [7], the adversarial loss is defined as the score produced by another neural network d_ψ, called critic, with some architecture parameterized by ψ. d_ψ learns to produce a large value when it takes a fake image generated by g_ϕ as input and produce a small value when it takes a real image as input. Conversely, g_ϕ is trained to deceive d_ψ, namely to force d_ψ to produce a small value when it takes a fake image as input. By increasing this loss with respect to ψ and decreasing it with respect to ϕ, g_ϕ is encouraged to generate images that are so realistic that d_ψ is unable to distinguish them from real face images. In addition to this loss, a penalty loss that encourages d_ψ to become Lipschitz continuous [16] is included in the training objective. These losses are collectively referred to as the image adversarial loss.

Attention Loss: With the architecture described above, if $\mathbf{A} = \mathbf{1}$, then $\hat{\mathbf{F}} = \mathbf{F}$. This means that since $\mathbf{F}$ is a real face image, the adversarial loss can be easily minimized with repect to ϕ when all the elements of $\mathbf{A}$ are 1, namely when g_ϕ always produces its input as it is. To avoid this situation, we need to guide the training so that as many elements of $\mathbf{A}$ as possible become zero. This corresponds to making the area of the input image to be converted as large as possible. In addition, the attention mask must be spatially smooth so that no discontinuous patterns will appear in the generated image. For these purposes, the norm and total variation of $\mathbf{A}$ are further included in the training objective. These losses are collectively referred to as the attention loss.

Conditional Expression Loss: In training, $\mathbf{y}$ is determined by extracting the AU intensities from a randomly selected image other than the input image. The generated image $\hat{\mathbf{F}}$ should be a face image with an expression consistent with that $\mathbf{y}$ (Fig. 2). Whether this is successful or not can be evaluated by the difference between $\mathbf{y}$ and the AU intensities extracted from $\hat{\mathbf{F}}$. To this end, another neural network r_ρ, an AU predictor, with some architecture parameterized by ρ is introduced. Since r_ρ must become a good AU predictor, the training objective should include a loss for the difference between $r_\rho(\mathbf{F})$ and $\mathbf{y}'$, where $\mathbf{y}'$ is the AU intensity vector extracted from $\mathbf{F}$ (using OpenFace). Also, for the purpose described above, the loss for the difference between $r_\rho(\hat{\mathbf{F}})$ and $\mathbf{y}$ are further included in the training objective. These losses are collectively referred to as the conditional expression loss. Although d_ψ and r_ρ can be defined as two separate networks, they can also be described as a multi-task network that branches into two heads near the end of the network.

Identity Loss: The identity of the face in $\hat{\mathbf{F}}$ will not necessarily be preserved if no constraints are imposed during training. As a way to address this, cycle-consistent training [17] is empirically known to be effective. The idea is to assume that converting an input image into a different domain and then converting back to the original domain must result in the original input image. In the current task, this can be induced by using a loss for the difference between $g_\phi(g_\phi(\mathbf{F}, \mathbf{y}), \mathbf{y}')$ and $\mathbf{F}$. This loss is referred to as the identity loss.

3.5. Facial Animation Synthesis from Speech

Once f_θ and g_ϕ have been trained, we can use them to animate a face image from speech (Fig. 3). First, we predict the AU intensity sequence $\hat{\mathbf{Y}} = [\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_N] = f_\theta(\mathbf{X})$ from the mel-spectrogram $\mathbf{X}$ of input speech. Given a still image input $\mathbf{F}$, we can generate its converted version $\hat{\mathbf{F}}_n = g_\phi(\mathbf{F}, \hat{\mathbf{y}}_n)$ by using $\hat{\mathbf{y}}_n$ for each n. Finally, we obtain a facial animation by concatenating all the generated images $\hat{\mathbf{F}}_1, \ldots, \hat{\mathbf{F}}_N$ in order.

4. Experiment

In this section, we present the results of objective and subjective evaluations. For the VoxCeleb2 dataset, we used the development set for training and the test set for testing, with the development set consisting of over a million videos and the test set consisting of over 36,000 videos. For the CelebA dataset, we used the first 162770 images for training and the remaining

Table 1: *MAEs and MSEs with 95% confidence intervals.*

Metric	Zero	MLP	RNN	CNN	DCNN
MAE	.3226±.0012	.2874±.0009	.2867±.0009	**.2819±.0009**	.2831±.0009
MSE	.3715±.0023	.2597±.0018	.2572±.0017	**.2516±.0017**	.2546 ± .0017

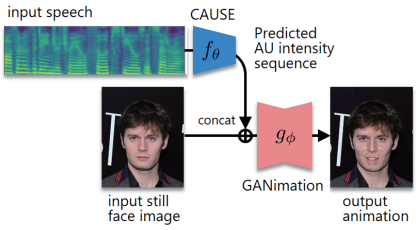

Figure 3: *Facial animation synthesis from speech*

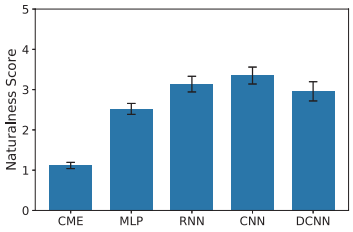

Figure 4: *Naturalness scores.*

images for testing.

First, we evaluated the performance of CAUSE for each architecture, where we used the audio part of each video in the test set as the input to CAUSE, and the AU intensity sequence extracted from the corresponding face-track sequence using OpenFace as the ground truth. For the objective performance measures, we used the mean absolute error (MAE) and mean square error (MSE). In addition to the predicted AU sequences obtained with MLP, RNN, CNN, and DCNN, we evaluated all-zero sequences ("Zero") as the minimum performance indicator. The results are shown in Table 1 along with 95% confidence intervals. As the results show, CNN performed the best, followed by DCNN, RNN, and MLP. We can draw some insights from these results. After visually checking some of the RNN predictions, we observed that there were many AU intensity peaks that had timing gaps between the predicted and ground truth sequences. Therefore, the use of the frame-wise error as the performance measure is probably the reason why the performance of RNN came out lower than expected. The fact that CNN worked slightly better than DCNN indicates that the cues for predicting the AU intensities of the speaker's face are concentrated only in a rather short segment of speech, near the current time. To put it a bit further, the speech segments distant from the current time are too noisy to be useful cues for AU prediction, and can rather confuse models that have the ability to capture long-term dependencies. CNN had a smaller receptive field than DCNN, allowing it to focus only on local segments, which may be the reason for its relatively higher performance. At the same time, the fact that CNN performed better than MLP suggests that prediction cues exist not only in the current frame but also in the neighboring frames.

We conducted a subjective test to evaluate the naturalness of the facial animations generated by the method described in 3.5 with each architecture. For comparison, we also implemented a version in which the AU intensity vectors are replaced with the emotional state distribution vectors extracted using the cross-modal transfer method [2]. We refer to this version as "cross-modal emotion (CME)". This version used the same CNN architecture as the one used in CAUSE and was trained using the FERPlus dataset [3]. Also, in this version, GANimation was trained using the emotional state distribution vectors as the conditioning variables instead. For this test, twenty-four subjects participated. Each participant was asked to watch each generated animation while listening to the corresponding input

Figure 5: *Examples of generated animations.*

speech and rate how well the face movement and expression in the animation matched the expression and mood of the speech by selecting 5: Excellent, 4: Good, 3: Fair, 2: Poor, or 1: Bad. We call this a naturalness score. The test was conducted online using Amazon Web Services, and each participant was asked to use a headphone in a quiet environment. The results are shown in Fig. 4. The results show that CAUSE was able to produce more natural animations than CME for all the architectures. This suggests the effectiveness of using the AU intensities as intermediate features in the task of generating facial animations from speech. In the comparisons of the four architectures, CNN, RNN, and DCNN were able to produce more natural animations than MLP. We believe that this is due to the fact that these three architectures have the ability to take account of the temporal dependencies in making AU sequence predictions, resulting in smoother and more natural animations. Fig. 5 shows two examples of generated animations when given the same speech. More examples can be found here[3].

5. Conclusion

In this paper, we proposed CAUSE, a method for estimating an AU intensity sequence from speech, and applied it to facial animation synthesis. For this, we built a neural network model that predicts an AU sequence from the mel-spectrogram of given speech and trained it using the speaking face-tracks in the VoxCeleb2 dataset and the AU sequences extracted by OpenFace as the ground truths. From the experimental results, we found that the CNN architecture was relatively better at handing the current tasks than the other architectures tested. We also found that using the AU intensities as intermediate features was found to be more effective than using the emotional state distributions for the task of generating facial animations from speech.

[3] http://www.kecl.ntt.co.jp/people/kameoka.hirokazu/Demos/cause/

6. References

[1] H. Fujisaki, "A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour," in *Vocal Physiology: Voice Production, Mechanisms and Functions*, O. Fujimura, Ed., pp. 347–355. Raven Press, New York, NY, USA, 1988.

[2] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," in *Proc. ACM Multimedia*, 2018.

[3] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proc. ICMI*, 2016.

[4] P. Ekman and W. V. Friesen, *Facial action coding system: a technique for the measurement of facial movement*, Consulting Psychologists Press, Palo Alto, CA, USA, 1978.

[5] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "GANimation: Anatomically-aware facial animation from a single image," in *Proc. ECCV*, 2018.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Adv. NIPS*, pp. 2672–2680, 2014.

[7] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. ICML*, pp. 214–223, 2017.

[8] M. Liao, S. Zhang, P. Wang, H. Zhu, X. Zuo, and R. Yang, "Speech2Video synthesis with 3d skeleton regularization and expressive body poses," in *Proc. ACCV*, pp. 308–323, 2020.

[9] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li, "MakeItTalk: Speaker-aware talking-head animation," *ACM Trans. Graph.*, vol. 39, no. 6, pp. 1–15, 2020.

[10] K. R. Prajwal, R. Mukhopadhyay, V. Namboodiri, and C. V. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proc. ACM MM*, pp. 484–492, 2020.

[11] T.-H. Oh, T. Dekel, C. Kim, I. Mosseri, W. T. Freeman, M. Rubinstein, and W. Matusik, "Speech2Face: Learning the face behind a voice," in *Proc. CVPR*, pp. 7539–7548, 2019.

[12] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *Proc. FG*, 2018.

[13] T. Baltrusaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," in *Proc. FG*, 2015.

[14] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech*, 2015.

[15] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. ICCV*, 2015.

[16] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein GANs," in *Adv. NIPS*, pp. 5769–5779, 2017.

[17] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-toimage translation using cycle-consistent adversarial networks," in *Proc. ICCV*, pp. 2223–2232, 2017.